Exploration of WordNet as a source of consumer health arthritis terms

Catherine K. Craven[ab], Olivier Bodenreider[b], Tony Tse[b]
University of Missouri, School of Medicine, Dept. of Health Management and Informatics
U.S. National Library of Medicine, National Institutes of Health, Dept. of Health and Human Services, USA

1. Introduction

Inclusion of patients as dynamic, proactive participants in their own healthcare is a relatively recent development in medicine. Not surprisingly, then, medical informatics efforts directly involving patients, their families and other non-experts as information users are newer areas of informatics focus. As a result, vocabularies for consumer health informatics applications that serve lay users have yet to be developed, explored or assessed to a large degree.

The need for vocabularies that nonexperts can understand and use for successful information retrieval is underscored by the numbers of consumers now seeking health information online: According to the May 2005 Health Information Online report by the Pew Internet & American Life Project, "Eight in ten internet users have looked online for health information....That translates into about 95 millions American adults (18+) who use the internet to find health information."[1]

The specialized language in which healthcare experts are educated is appropriate for communication amongst their peers. Many controlled terminologies for representing specialized health expert language have been created, explored, assessed and harnessed for information representation, information retrieval and text mining purposes, e.g. MeSH, ICD-9CM, SNOMED CT and the many other vocabularies incorporated into the National Library of Medicine's (NLM) Unified Medical Language System's® (UMLS®) Metathesaurus®. However, the specialized language used by experts amongst themselves and for their own purposes is not always ideal or appropriate for use in communication with patients. The language of experts can add noise to health communication channels whether those channels be healthcare provider to patient face-to-face, print or online media source (e.g. health articles in Newsweek online) to health consumer, or online consumer health informatics information resources (e.g. MedlinePlus) to health consumer. "Patrick et al. [2] have called the "consumer vocabulary problem" a fundamental issue in health information provision…the mismatch between terms used by healthcare professionals and those used by consumers who receive their services."[2] Thus, vocabularies for communication with and for information retrieval by non-experts need development.

Tse has identified vocabulary categories by user, including the Professional Medical Vocbulary (PMV), the Medical Mediator Vocabulary (MMV) and the Consumer Medical Vocabulary (CMV).[3] The vocabularies included in the UMLS are PMVs. Tse defines MMVs as terms "extracted from health-related documents authored by professional information intermediaries (e.g. health communicators, journalists, and librarians)."[3] Other studies have collected lay terms through a variety of means including search log queries from consumer health Web sites, user e-mail, and chat forums. Because of time and methodological constraints involved in collecting actual lay user terms, in this preliminary study we use terms from a medical mediator document corpus to represent lay terms.

However, we justify these MMV terms as an acceptable surrogate for lay terms because generally health communicators attempt to write in a straightforward, concise manner that is easily understood by a broad lay audience and, accordingly, try to use lay language where possible. A case can be made that an MMV, which Tse hypothesizes is "a natural bridge between CMV and PMV"[3], is in fact a desirable basis for a CMV; although medical mediators usually keep language simple, which assists health consumers new to a topic and/or of a lower literacy level, medical mediators do cover a broad range of topics pertinent to their topics including new drug and research findings and even cutting-edge techniques still under development, the language and concepts – the "terms" -- of which serve to educate even the more knowledgeable health information seekers and those with higher literacy levels.

In this study, we explore the lexical database WordNet as a possible CMV. In other studies, WordNet is explored for many purposes in research areas including computational linguistics, natural language processing and artificial intelligence. A few studies examine WordNet for specialized biomedical domains. Burgun and Bodenreider (2001) compared terms, concepts and semantic classes in WordNet and the UMLS in both a general class "ANIMAL" and the domain-specific class, "HEALTH DISORDER" [4] and (2002) characterized the definitions of anatomical concepts in WordNet and a medical dictionary. [5] Terms for four broad categories – phenotype, molecular function, biological process and cellular component were mapped to WordNet in an evaluation of it as a source of lay knowledge for potential use in a consumer health system on genetic diseases.[5,6] These studies encouraged us to probe WordNet as a source of terms that could be used as the basis of a consumer health

vocabulary on arthritis, the leading cause of disability in the United States, “limiting the activities of more than 16 million adults.”[7]

2. Materials

Consumer health document corpus on arthritis

One consumer health article from the WebMD site was selected for manual term identification and mapping. The approximately 31,000-word main document corpus from which terms were automatically extracted and then mapped comprises 17 arthritis-related publications with a wide distribution/readership and that target a broad consumer audience. The corpus articles cover several MMV genres: informational, persuasive and human interest.[8] Ads for health-related products, which fall into the commercial MMV genre, were excluded for brevity. Sources for the articles include the *New York Times*; *Los Angeles Times*; *St. Louis Post-Dispatch;* publications linked to the NLM’s MedlinePlus consumer health Web site from the American Academy of Orthopedic Surgeons, the Arthritis Foundation, the Centers for Disease Control, the Cleveland Clinic, National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), *Arthritis Today*, and United Press International; publications linked to WebMD from *The Arthritis Sourcebook*, NIAMS; *Newsweek; U.S. News & World Report.*

WordNet

WordNet®, developed since 1985 at the Cognitive Science laboratory at Princeton University under Professor George A. Miller’s direction, is an online lexical database, a “reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.”[9] Relationships depend on the part of speech. WordNet nouns, which are normalized to the singular, are of primary interest as opposed to the parts of speech “because the vast majority of biomedical terms are noun phrases in which the head noun is modified by either an adjective, another noun or a prepositional phrase.”[4] Specifiable relationships in WordNet among the noun synonym sets or “synsets” – those synonyms that convey the same sense or meaning of a given noun -- include the following main relationships among others: synonyms, coordinate terms (i.e. nouns or verbs that have the same hypernym), hypernyms, hyponyms, holonyms and meronyms. Pronouns are not included. Limited proper names are specified in the database. Currently, the database (v. 2.1) includes 117,097 unique noun strings, 81,426 noun synsets and about a 117,000 total synsets from all four parts of speech.[9]

Unified Medical Language System’s Metathesaurus®

Developed and maintained by the NLM since 1990, the purpose of the UMLS “is to facilitate the development of computer systems that behave as if they ‘understand’ the meaning of the language of biomedicine and health”[10] The UMLS consists of three Knowledge Sources (databases): the Metathesaurus, the Semantic Network and the Specialist Lexicon as well as software programs to manipulate them. This database and tool set are used by system developers and informatics researchers “in building or enhancing electronic information systems that create, process, retrieve, integrate, and/or aggregate biomedical and health data and information” for application “in systems that perform a range of functions involving one or more types of information, e.g., patient records, scientific literature, guidelines, public health data.”[10] The Metathesaurus is a repository of over one million biomedical concepts and five million concept names from more than 100 controlled professional medical vocabularies and classifications.

Structure among the terms is provided in several ways. Information about each term’s hierarchical position and other inter-term relationships from its source vocabulary is preserved. The terms from the component vocabularies in the Metathesaurus are further organized by concept or meaning. Varying names for the same concept (synonyms, lexical variants, and translations) are linked together via a concept unique identifier or CUI. Each Metathesaurus concept has assigned to it, from the Semantic Network, one of 135 “semantic types” -- broad categories. The Semantic Network specifies also which of 54 relationships are possible between each semantic type.

PhrasEx

The NLM-developed PhrasEx is a program for shallow syntactic parsing to extract noun phrases from electronic text. PhrasEX interacts with and relies on three other components, the Specialist minimal commitment parser, UMLS Specialist Lexicon and the MedPost-SKR stochastic tagger. PhrasEx refers to "the syntactic structure provided by the Specialist Minimal commitment parser [10], which relies on the Specialist Lexicon [11] as well as the MedPost-SKR stochastic tagger to resolve part-of-speech ambiguity [12]."[11]

PhrasEx drops determiners and pronouns and "tokenizes its input using non-alphanumeric characters as token separators and preserves the case."[11] PhrasEx outputs three types of noun phrases: "**simp**", "**macro**" and "**mega.**" For example, in the sentence "Your hip is a ball and socket joint, formed by the upper end of the femur, the ball, and a part of the pelvis called the acetabulum, the socket," PhrasEx extracted multiple phrases. Ten **simp** phrases were extracted: "socket", "acetabulum", "pelvis", "part", "ball", "femur", "upper end", "formed", "ball and socket joint", "hip." Two **macro** phrases "part of pelvis" and "formed by upper end of femur" were extracted. Two **mega** phrases "hip ball and socket joint formed by upper end of femur ball and part of pelvis" and "acetabulum socket" were extracted. **Simp** phrases then, are those with a head noun. **Macro** phrases have prepositional modification to the right, and although the "first preposition is unconstrained, the rest must be *of*."[11] **Mega** phrases include "all the content words in the sentence to the left and right of a finite verb."[11]

3. Methods

Manual extraction

The first step was an exploration of WordNet 2.1 through a small manual "extraction" of medically related noun phrases from the document identified for this purpose. Both one-word and compound noun phrases were highlighted in the document. Compounds here are the simplest individual lexical unit where the meaning of the noun phrase would be lost if additional modifiers – adjectives or nouns – were removed. A few non-medical nouns such as "typing" and "needlework" were left unhighlighted; due to the conciseness of the test document, however, nearly every noun phrase was selected. Duplicate noun phrases were removed and phrases were uninflected (i.e. normalized to be singular). No spelling normalization was required. Each identified noun phrase was entered manually into WordNet. We recorded whether or not we found a noun and accompanying "synset" in WordNet that matched the normalized extracted string exactly. In cases of polysemy, where more than one synset matched the noun string, we disambiguated by recording which synset was the appropriate one for the phrase in the context of the document. For each extracted noun phrase that did not have an exact string match in WordNet, we then did an additional search for alternative-string synonyms and ranked any synonym that we found as an exact synonym, a close synonym or a partial synonym.

Automated extraction

The main effort was a largely automated effort to mimic and expand the manual process above. The 17-article main corpus was formatted into a single .txt file and input into PhrasEx, which extracted noun phrases into the simp, macro and mega formats. We then removed the phrase duplicates that were caused by PhrasEx from the 15,107 phrases it generated in the extraction process. There were many duplicates because the same phrase can appear several times in the PhrasEx output under different categories: simp, macro and mega. The issue is that this artificially increases the frequencies for tokens. In a Perl program to cull actual tokens, we adopted the precedence that: simp > macro > mega. In other words, only mega phrases that are not also simp or macro were recorded. We did not normalize for plurals at this point, which does skew the phrase frequencies but did not affect our mappings.

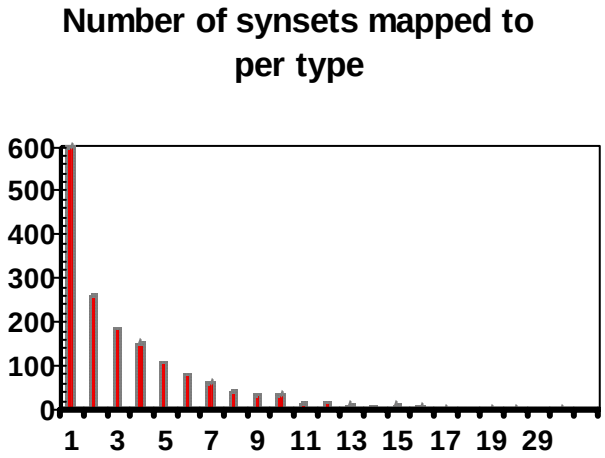
Mapping

We mapped the extracted phrases to WordNet using the standard function *wn* integrated into the WordNet interface. We also wanted to map the extracted phrases to the UMLS as a benchmark for biomedical terms. We mapped the phrases to the UMLS Metathesaurus, version 2005AA using a Knowledge Source Server API. Our mapping strategy was to attempt exact matches followed by normalized matches if exact match fails.

Disambiguation

The phrases that do map to something in WordNet can be mapped to multiple synsets, only one of which is associated with the correct biomedical meaning, or they can map to one synset where the string is correct but the meaning is wrong [Figure 1].

Figure 1 -- Synsets mappings with potential for ambiguity



To initiate disambiguation of the synsets, we followed a technique developed in 2001 by Burgun and Bodenreider [4] for establishing semantic categories of biomedical interest in WordNet. We create a semantic category in WordNet by searching up the term hierarchy for, as an example, several anatomical parts such as "joint," "bone," "foot," "nerve," and "hip," and finding their least common hypernym, which for anatomical parts happens to be "body part." The hyponyms below it – and for categories such as anatomy, meronyms, too -- are the synsets to be included in the new semantic category.

We created two semantic categories in WordNet, one for anatomy and one for diseases, and then supplemented them with several semi-complete semantic categories for procedures, drugs, medical devices, micro-organisms and chemicals, all of which will need further development. We then wrote a short Perl program to identify the synsets belonging to these categories. This allows us to distinguish between biomedical and general language synsets.

4. Results

Manual effort

For the manual phrase extraction, 148 unique phrases resulted. WordNet yielded exact string matches with the correct meaning for 96 (65%) of the noun phrases. Fifty-two noun phrases were not found. One of those unfound noun phrases is a drug category acronym, "DMARD" and another unfound phrase is the drug "sulfasalazine." The other 50 of those non-covered noun phrases were compounds. Most were two-word compounds in which the meaning of the noun phrase would be lost if the modifying noun or adjective were removed. A few of the compounds were longer phrases, such as "biologically engineered drug." The lack of semantic linking between adjectives and nouns is an acknowledged WordNet drawback.[12]

We then re-queried WordNet for alternate-string synonyms for each of the 50 noun compounds without exact string matches. We found 18 additional WordNet nouns that were either exact, close or partial synonyms. If we add only the exact and close synonyms to the WordNet coverage of our identified noun phrases, that figure increases 108 matches out of 148, for a coverage rate of about 73%.

Automated effort

After culling actual tokens from the 15,107 phrases generated by PhrasEx in the extraction process, the result was 13,770 actual tokens and 8,916 types or unique noun phrases. We found that 1,634 out of the 8,916 types mapped to at least one noun synset in WordNet for a mapping rate of about 18 % [Table 1]. Of the three PhrasEx phrase categories, the **simp** types accounted for most of the type mappings that did occur, which is not surprising; **simp** phrases are the closest strings to the nouns that populate WordNet, the majority of which are one-word nouns, along with some common compounds.

Table 1 -- Mapping of extracted phrases to WordNet*

(*WordNet 2.0)	No WordNet Mapping	Some WordNet Mapping
Total tokens	8,581 **(62%)**	5,189 **(38%)**
Simp	4,003	5,140
Macro	1,708	4
Mega	2,870	45
Total types	7,282 **(82%)**	1,634 **(18%)**
Simp	3,074	1,604
Macro	1,617	4
Mega	2,591	26

We saw reassuringly, that many important, common arthritis phrases such as *rheumatoid arthritis*, *osteoarthritis*, *inflammation*, *joint* and *cartilage* all mapped to something in WordNet. Of concern, however, is that included in that not-mapped list are important common arthritis phrases such as *bextra*, *fibromyalgia*, *rheumatic disease*, *joint pain*, and *total joint replacement*.

In our benchmark mapping of the extracted phrases to the UMLS Metathesaurus, we found 2,120 types, or about 24% of the types, with some mapping in the UMLS [Table 2].

Table 2 – Mapping of extracted phrases to the UMLS

(*UMLS2005AA)	No Metathesaurus Mapping	Some Metathesaurus Mapping
Total tokens	7786 **(57%)**	5984 **(43%)**
simp	3,501	5,642
macro	1,571	141
mega	2,714	201
Total types	6796 **(76%)**	2120 **(24%)**
simp	2,826	1,852
macro	1,503	118
mega	2,467	150

Disambiguation

Out of the 1,634 types that mapped to something in WordNet, 1,256 (77%) do not correspond to our semantic categories; 378 (23%) types map to (at least) one synset and do correspond to (at least) one of our categories. Nine types correspond to several categories. For example, “joint” corresponds with both the category “anatomy” and the category “drug” because “joint” is a synomyn for marijuana.

One hundred ninety-three of the 378 types mapped to exactly one WordNet synset, of which one corresponded to two categories. For the other 237 cases, which had mappings to multiple synsets, the semantic categories do identify which of the mapped-to synsets are biomedical. In 8 cases, however, several of the synsets correspond to biomedical categories and further information is required to fully disambiguate the original phrase.

5. Discussion

More types mapped to something in the Metathesaurus (24%) than did types map to something in WordNet (18%), which provides some confirmation that the types extracted are medically related. That only 18% of phrases mapped to something in the automated effort in WordNet is far from the promising 65% exact match coverage for the manual mini-effort. Our limited study suggests that WordNet might not provide enough coverage of lay medical terms to be useful as a consumer health vocabulary source. However, these findings are to be confirmed by an expanded study.

6. Future Work

We'd like to expand the corpus substantially, and at the same time increase the variety of MMV sources, genres and targeted audiences. We'll normalize for plurals up front in the next phase. And we'll increase our disambiguation efforts by verifying and completing our current semantic categories and identify more. In addition, we'll deal with the "silence," those phrases that map to nothing in WordNet such as *bextra*, *fibromyalgia*, *rheumatic disease*, *joint pain*, and *total joint replacement*, by automating an alternate-string synonym search process. We'll accomplish this by joining the file of phrase mappings to WordNet and the file of phrase mappings to the UMLS. For any phrase that does *not* map to WordNet but *does* map to the Metathesaurus, we'll identify the associated UMLS concept unique identifier, the CUI, and all of the alternate-string synonyms associated with it. We'll then take those alternate-strings and remap them to WordNet to see if we can find matches in WordNet. This approach would, for example, automatically identify the synset *arthralgia* in WordNet as a mapping for the phrase *joint pain*, through the synonymy recorded in the Metathesaurus between the two terms (C0003862).

7. References

[1] Fox, S. Health Information Online. Pew Internet & American Life Project. 2005 May 17; 22; Available from: http://www.pewinternet.org/pdfs/PIP_Healthtopics_May05.pdf.

[2] Smith CA, Stavri PZ. Consumer health vocabulary. In: Lewis D, Eysenbach G, Kukafka R, Stavri PZ and Jimison HB, editors. Consumer health informatics: informing consumers and improving health care. New York: Springer. 2005.

[3] Tse AY and Soergel D. Procedures for mapping vocabularies from non-professional discourse a case study: "consumer medical vocabulary." In Proceedings for the American Society for Information Science and Technology Annual Meeting; 2003: Long Beach, Calif.: American Society for Information Science and Technology. p. 174-83.

[4] Burgun A, Bodenreider O. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In: Proceedings of the North American Chapter of the Association for Computational Linguistics 2001 Workshop WordNet and Other Lexical Resources, Applications, Extensions and Customizations; 2001: Pittsburgh, Pennsylvania; 2001. p. 77-82.

[5] Burgun A, Bodenreider O. characterizing the definitions of anatomical concepts in WordNet and specialized sources. In: Proceedings of the First Global WordNet Conference; 2002: Mysore, India: Central Institute for Indian Languages; 2002. p. 223-230.

[6] Bodenreider O, Burgun A, Mitchell JA. Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. In: Baud M, Fieschi M, Le Beux P, Ruch P, editors. Studies in health technology and informatics. Volume 95. The New Navigators: from Professionals to Patients**.** Proceedings of MIE2003. Amsterdam: IOS Press; 2003. p. 379-84.

[7] Centers for Disease Control. National Center for Chronic Disease Prevention and Health Promotion. Arthritis. Atlanta, Georgia. [updated 2005 June 28; accessed 2005 July 18]. Available from: http://www.cdc.gov/arthritis/index.htm

[8] Tse AY. Identifying and characterizing a "consumer medical vocabulary." [dissertation]. College Park (MD): University of Maryland; 2003.

[9] http://wordnet.princeton.edu/

[10] http://www.nlm.nih.gov/research/umls/about_umls.html

[11] Srinivasan S, Rindflesch TC, Hole WT, Aronson AR, Mork JG. Finding UMLS Metathesaurus concepts in MEDLINE. In: Proceedings of the AMIA Annual Symposium; 2002: San Antonio, Texas: American Medical Informatics Association. p. 727-731.
[12] Fellbaum, C, editor. WordNet: an electronic lexical database. Cambridge, Mass.: The MIT Press; 1998.